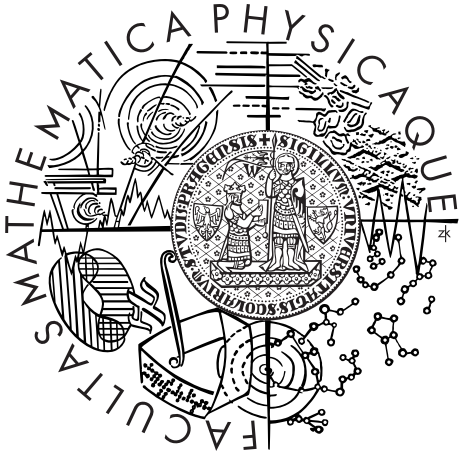
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Katarína Vlčková

Hardyho-Weinbergova rovnováha

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Karel Zvára CSc.

Studijní program: Matematika

Studijní obor: Obecná matematika

Praha 2012

Na tomto mieste by som sa chcela poďakovať vedúcemu mojej bakalárskej práce doc. RNDr. Karlovi Zvárovi CSc. za pomoc, ochotu a trpezlivosť, ktorou sa mi vždy venoval.

Prehlasujem, že som túto bakalársku prácu vypracovala samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, hlavne skutočnosť, že Univerzita Karlova v Prahe má právo na uzavretie licenčnej zmluvy o užití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V ..... dňa ..... Podpis autora

Názov práce: Hardyho-Weinbergova rovnováha

Autor: Katarína Vlčková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: doc. RNDr. Karel Zvára CSc. Katedra pravděpodobnosti a matematické statistiky MFF UK

Abstrakt: Táto práca popisuje štatistické testy, používané pri riešení otázky, či je populácia v Hardyho-Weinbergovej rovnováhe. Konkrétne popisuje exaktný test, χ^2 test, χ^2 test s opravou na spojitosť, modifikovaný χ^2 test, modifikovaný χ^2 test s opravou na spojitosť a test pomerom vierohodnosti. Tieto testy riešia otázku, či sa daný náhodný výber dá popísať trinomickým rozdelením, ktorého pravdepodobnosti sú tvaru $p_{AA} = \theta^2$, $p_{Aa} = 2\theta(1-\theta)$, $p_{aa} = (1-\theta)^2$. Niektoré testy sú výhodnejšie na použitie než iné. Pri simulácii pre veľkosť populácie 100 jedincov sa najvhodnejším javí modifikovaný χ^2 test. Odhady sily testov pre χ^2 test a test pomerom vierohodnosti tiež vykazujú veľmi dobré výsledky v porovnaní s ostatnými testami, ale čo sa týka odhadu hladiny testov, javia sa v niektorých prípadoch antikonzervatívne.

Kľúčové slová: Hardyho-Weinbergová rovnováha, multinomické rozdelenie, χ^2 test, test pomerom vierohodnosti, exaktný test

Title: Hardy-Weinberg equilibrium

Author: Katarína Vlčková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Karel Zvára CSc. Department of Probability and Mathematical Statistics

Abstract: In this paper, we describe various tests used to determine deviations from the Hardy-Weinberg equilibrium. The tests described are: the exact test, the χ^2 test with and without continuity correction, the conditional χ^2 test with and without continuity correction and the likelihood ratio test. These tests explore the question whether a random sample has trinomic distribution with probabilities $p_{AA} = \theta^2$, $p_{Aa} = 2\theta(1-\theta)$, $p_{aa} = (1-\theta)^2$.
In this work, we simulate data of sample size 100 and we estimate the probability of type I error and the power of the tests. In this case, we get the best results with conditional χ^2 test. The estimate of the power of the likelihood ratio test and the χ^2 test is one of the highest of all. On the other hand, these two test are anticonservative in some cases .

Keywords: Hardy-Weinberg equilibrium, multinomic distribution, χ^2 test, likelihood ratio test, exact test

Obsah

Úvod **6**

1 Multinomické rozdelenie **7**

2 Exaktný test **10**

3 Testy χ^2 pri neznámych parametroch **12**

4 Modifikovaný χ^2 test **14**

5 Test pomerom vierohodnosti **16**

6 Porovnávanie testov **21**
6.1 Grafické znázornenie niektorých testov 21
6.2 Hladina testov 22
6.3 Sila testov 23

Záver **30**

Zoznam použitej literatúry **31**

Zoznam tabuliek **32**

Zoznam obrázkov **33**

Zoznam príloh **34**

Úvod

Hardyho-Weinbergova rovnováha je princíp využívaný v populačnej genetike. Je to ideálny stav, ktorý slúži ako základ k ziťovaniu a meraniu zmien v populácii. Hardyho-Weinbergov zákon bol odvodený na sebe nezávisle matematikom G. H. Hardym a fyzikom Wilhelmom Weinbergom v roku 1908.
Z biologického hladiska tento princíp tvrdí, že pravdepodobnosti výskytu aliel a genotypov zostávajú konštantné z generácie na generáciu. Rovnováha môže byť porušená ak:

- párenie nie je náhodné,
- nastávajú mutácie,
- veľkosť populácie je obmedzená,
- nastáva tok génov.

V mojej práci sa zaoberám prípadom diploidných jedincov (každý jedinec má dve alely) a génmi, ktoré sú určované dvoma alelami.
Predmetom tejto práce je popis štatistických testov, ktoré sa používajú pri overovaní, či sa pravdepodobnosti výskytu genotypov v populácii líšia od pravdepodobností v prípade splnenia Hardyho-Weinbergovej rovnováhy. Prvé kapitoly takéto testy – konkrétne exaktný test, χ^2 test spolu s rôznymi modifikáciami a test pomerom vierohodnosti – popisujú a overujú oprávnenosť ich použitia v danej situácii. Posledná kapitola je venovaná odhadovaniu hladiny a sily testov a grafickému znázorneniu niektorých uvedených testov.

1. Multinomické rozdelenie

Definícia 1.1. *Uvažujme n rôznych nezávislých pokusov. Pri každom pokuse nastane práve jeden z k javov a pravdepodobnosť, že nastane jav i je p_i, $0 < p_i < 1$, $i = 1, ..., k$. Označme X_i počet pokusov, v ktorých nastal jav i. Potom združené rozdelenie náhodného vektoru $(X_1, X_2, ..., X_k)'$ nazývame multinomické. Označujeme ho $M(n, p_1, p_2, ..., p_k)$ a platí:*

$$P(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

pre $x_i = 0, ..., n, i = 1, ..., k$ a $x_1 + ... + x_k = n$

Ďalej zaveďme označenie $\boldsymbol{X} = (X_1, X_2, ..., X_k)'$ a $\boldsymbol{p} = (p_1, p_2, ..., p_k)'$.

Veta 1.1. *Predpokladajme, že $\boldsymbol{X}$ má multinomické rozdelenie $\boldsymbol{X} \sim M(n, \boldsymbol{p})$. Zvoľme $1 < l \leq k$. Platí:*
a) marginálne rozdelenie veličín $X_l, ..., X_k$ je dané vzorcom

$$P(X_l = x_l, ..., X_k = x_k) = \frac{n!}{x_l!...x_k!(n - x_l - ... - x_k)!} p_l^{x_l} ... p_k^{x_k} (1 - p_l - ... - p_k)^{n - x_l - ... - x_k}$$

b) podmienené rozdelenie veličín $X_1, ..., X_{l-1}$ za podmienky $X_l = x_l, ..., X_k = x_k$ je dané vzorcom

$$P(X_1 = x_1, ..., X_{l-1} = x_{l-1} \mid X_l = x_l, ..., X_k = x_k) = \frac{(n - x_l - ... - x_k)!}{x_1!...x_{l-1}!} \prod_{i=1}^{l-1} \left(\frac{p_i}{1 - p_l - ... - p_k} \right)^{x_i},$$

kde $x_i = 0, 1, ..., n - x_l - ... - x_k$, $i = 1, ..., l - 1$ a $x_1 + ... + x_k = n$.

Dôkaz. a) Použijeme multinomický rozvoj.

$$\begin{aligned} P(X_l = x_l, ..., X_k = x_k) &= \sum_{x_1, ..., x_{l-1}} \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k} \\ &= \frac{n!}{x_l!...x_k!(n - x_l - ... - x_k)!} p_l^{x_l} ... p_k^{x_k} \sum_{x_1, ..., x_{l-1}} \frac{(n - x_l - ... - x_k)!}{x_1!...x_{l-1}!} p_1^{x_1} ... p_{l-1}^{x_{l-1}} \\ &= \frac{n!}{x_l!...x_k!(n - x_l - ... - x_k)!} p_l^{x_l} ... p_k^{x_k} (p_1 + ... + p_{l-1})^{n - x_l - ... - x_k} \\ &= \frac{n!}{x_l!...x_k!(n - x_l - ... - x_k)!} p_l^{x_l} ... p_k^{x_k} (1 - p_l - ... - p_k)^{n - x_l - ... - x_k} \end{aligned}$$

Pričom v sumách sčítavame cez všetky také $x_1, ..., x_{l-1}$, pre ktoré platí $x_1 + ... + x_{l-1} = n - x_l - ... - x_k$.
b) z Bayesovho vzorca plynie

$$P(X_1 = x_1, ..., X_{l-1} = x_{l-1} \mid X_l = x_l, ...X_k = x_k) = \frac{P(X_1 = x_1, ..., X_{l-1} = x_{l-1} \cap X_l = x_l, ..., X_k = x_k)}{P(X_l = x_l, ...X_k = x_k)}$$

Z časti a) teda plynie:

$$\begin{aligned}
&P(X_1 = x_1, ..., X_{l-1} = x_{l-1} \mid X_l = x_l, ...X_k = x_k) \\
&\quad = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k} \frac{x_l!...x_k!}{n!} \frac{(n - x_l - ... - x_k)!}{p_l^{x_l} ... p_k^{x_k} (1 - p_l - ... - p_k)^{n - x_l - ... - x_k}} \\
&\quad = \frac{(n - x_l - ... - x_k)!}{x_1!...x_{l-1}!} \prod_{i=1}^{l-1} \left(\frac{p_i}{1 - p_l - ... - p_k} \right)^{x_i}
\end{aligned}$$

□

Veta 1.2. *Nech $\boldsymbol{X}$ je náhodný vektor s multinomickým rozdelením $\boldsymbol{X} \sim M(n, \boldsymbol{p})$. Potom platia nasledujúce rovnosti:*

$$\begin{aligned}
&EX_i = np_i, && 1 \leq i \leq k \\
&varX_i = np_i(1 - p_i), && 1 \leq i \leq k \\
&cov(X_i, X_j) = -np_ip_j, && 1 \leq i, j \leq k, i \neq j
\end{aligned}$$

Dôkaz. Zvoľme ľubovolné $i, j = 1, ..., k$
Z vety 1.1 vyplýva, že $P(X_i = x_i) = \frac{n!}{x_i!(n-x_i)!} p_i^{x_i} (1 - p_i)^{n - x_i}$, $i = 1, ..., k$.
Z toho výpočtom zistíme:

$$\begin{aligned}
EX_i &= \sum_{l=0}^{n} l \frac{n!}{l!(n-l)!} p_i^l (1 - p_i)^{n-l} \\
&= np_i \sum_{l=1}^{n} \frac{(n-1)!}{(l-1)!((n-1) - (l-1))!} p_i^{l-1} (1 - p_i)^{(n-1)-(l-1)} \\
&= np_i
\end{aligned}$$

z binomického rozvoja.

$$\begin{aligned}
EX_i(X_i - 1) &= \sum_{l=0}^{n} l(l-1) \frac{n!}{l!(n-l)!} p_i^l (1 - p_i)^{n-l} \\
&= n(n-1)p_i^2 \sum_{l=2}^{n} \frac{(n-2)!}{(l-2)!((n-2) - (l-2))!} p_i^{l-2} (1 - p_i)^{(n-2)-(l-2)} \\
&= n(n-1)p_i^2
\end{aligned}$$

$$\begin{aligned}
&EX_i^2 = EX_i(X_i - 1) + EX_i = n(n-1)p_i^2 + np_i \\
&varX_i = n(n-1)p_i^2 + np_i - n^2 p_i^2 = np_i(1 - p_i)
\end{aligned}$$

Definujme $X_{ij} := X_i + X_j$. Potom platí:

$$var(X_{ij}) = n(p_i + p_j)(1 - p_i - p_j)$$

$$\begin{aligned}
var(X_{ij}) &= var(X_i + X_j) = var(X_i) + var(X_j) + 2cov(X_i, X_j) \\
&= np_i(1 - p_i) + np_j(1 - p_j) + 2cov(X_i, X_j)
\end{aligned}$$

Z toho vyplýva:

$$2cov(X_i, X_j) = np_i - np_i^2 - 2np_ip_j - np_j - p_j^2 - np_i + np_i^2 - np_j + np_j^2$$
$$cov(X_i, X_j) = -np_ip_j$$

□

Uvažujme výber z populácie s genotypmi AA, Aa, aa. Označme počet jedincov s genotypom AA ako X_{AA}, s genotypom Aa ako X_{Aa} a s genotypom aa ako X_{aa}. Predpokladajme, že $X = (X_{AA}, X_{Aa}, X_{aa})'$ je náhodný vektor s trinomickým rozdelením a s pravdepodobnosťami $p = (p_{AA}, p_{Aa}, p_{aa})'$. Potom z predchádzajúcich viet plynie:

$$\begin{aligned}
EX_{AA} &= np_{AA} \\
EX_{Aa} &= np_{Aa} \\
EX_{aa} &= np_{aa} \\
varX_{AA} &= np_{AA}(1 - p_{AA}) \\
varX_{Aa} &= np_{Aa}(1 - p_{Aa}) \\
varX_{aa} &= np_{aa}(1 - p_{aa}) \\
cov(X_{AA}, X_{Aa}) &= -np_{AA}p_{Aa} \\
cov(X_{AA}, X_{aa}) &= -np_{AA}p_{aa} \\
cov(X_{Aa}, X_{aa}) &= -np_{Aa}p_{aa}
\end{aligned}$$

Počet aliel A a a môžeme popísať náhodnými veličinami X_A, X_a, nech X_A udáva počet aliel A a X_a udáva počet aliel a. V prípade nezávislého združovania, tj. aj v prípade, že populácia spĺňuje Hardyho-Weinbergovú rovnováhu môžeme rozdelenie týchto náhodných veličín popísať binomickým rozdelením s pravdepodobnosťami $p_A = \theta$ a $p_a = 1 - \theta$, tj.:

$$P(X_A = x_A, X_a = x_a) = \frac{2n!}{x_A!x_a!}\theta^{x_A}(1 - \theta)^{x_a},$$

kde $2n$ je počet génov vo výbere z populácie a $x_A + x_a = 2n$. Populácia je v Hardyho-Weinbergovej rovnováhe ak $p_{AA} = p_A^2 = \theta^2$, $p_{Aa} = 2p_Ap_a = 2\theta(1 - \theta)$, $p_{aa} = p_a^2(1 - \theta)^2$.
V ďalších kapitolách budeme testovať, či je daná populácia v Hardyho-Weinbergovej rovnováhe, tj. testujeme nulovú hypotézu $H_0 : p_{AA} = \theta^2, p_{Aa} = 2\theta(1 - \theta), p_{aa} = (1 - \theta)^2$, voči alternatívnej hypotéze, ktorá je definovaná nasledovne: H_1 : aspoň jedna z nasledovných nerovností je splnená $p_{AA} \neq \theta^2, p_{Aa} \neq 2\theta(1 - \theta), p_{aa} \neq (1 - \theta)^2$, resp. nulovú hypotézu $H_0 : p_{Aa}^2 = 4p_{AA}p_{aa}$ voči alternatívnej hypotéze $p_{Aa}^2 \neq 4p_{AA}p_{aa}$. V druhom prípade vieme testovať H_0 voči jednostranným alternatívam: $H_1 : p_{Aa}^2 > 4p_{AA}p_{aa}$, resp. $p_{Aa}^2 < 4p_{AA}p_{aa}$.

2. Exaktný test

Majme populáciu v Hardyho-Weinbergovej rovnováhe. V kapitole 1 sme odvodili pravdepodobnosť, že alela A sa za n pokusov vyskytne vo výbere x_A-krát:

$$P(X_A = x_A, X_a = x_a) = \frac{2n!}{x_A! x_a!}\theta^{x_A}(1-\theta)^{x_a},$$

Podmienená pravdepodobnosť javu $X_{AA} = x_{AA}$, $X_{Aa} = x_{Aa}$, $X_{aa} = x_{aa}$ za podmienky $X_A = x_A$, $X_a = x_a$ bude:

$$P(X_{AA} = x_{AA}, X_{Aa} = x_{Aa}, X_{aa} = x_{aa} \mid X_A = x_A, X_a = x_a) \tag{2.1}$$

$$= \frac{n!}{x_{AA}! x_{Aa}! x_{aa}!} p_{AA}^{x_{AA}} p_{Aa}^{x_{Aa}} p_{aa}^{x_{aa}} \frac{x_A! xa!}{2n!} \frac{1}{p_A^{x_A} p_a^{x_a}} \tag{2.2}$$

$$= \frac{n!}{x_{AA}! x_{Aa}! x_{aa}!} 2^{x_{Aa}} \frac{x_A! xa!}{2n!}$$

Bez ujmy na obecnosti nech $x_A < x_a$. Potom zo vzťahu $x_A = 2x_{AA} + x_{Aa}$ vyplýva, že x_A je párne práve vtedy, keď x_{Aa} je párne. Túto informáciu využijeme neskôr. Potrebujeme zistiť P-hodnotu testu.

a) Jednostranné testy

P-hodnotu určujeme ako súčet všetkých pravdepodobností (2.1) vhodných situácií za predpokladu, že x_A je pevné. V prípade, že $H_1 : p_{Aa}^2(\theta) < 4p_{AA}(\theta)p_{aa}(\theta)$, získame P-hodnotu ako $P_D = P(X_{Aa} \leq x_{Aa} \mid X_A = x_A, X_a = x_a)$. Túto hodnotu môžeme získať ako súčet pravdepodobností (2.1), kde počet aliel A a a zostáva stály, počet heterozygotov j bude menší alebo rovný ako x_{Aa} a počet homozygotov AA bude splňovať $x_A = 2x_{AA} + j$. Ďalej pri výpočtoch musíme brať do úvahy párnosť, resp. nepárnosť x_A.

Podobný postup sa používa pri $H_1 : p_{Aa}^2(\theta) > 4p_{AA}(\theta)p_{aa}(\theta)$, pre P-hodnotu platí $P_H = P(X_{Aa} \geq x_{Aa} \mid X_A = x_A, X_a = x_a)$. Podobne ako v predchádzajúcej situácii P_H získame ako súčet pravdepodobností (2.1), kde počet aliel A a a zostáva stály, počet heterozygotov j bude väčší alebo rovný ako x_{Aa}, počty homozygotov budú splňovať podmienku $x_A = 2x_{AA} + j$.

b) Dvojstranný test

P-hodnotu môžeme definovať dvoma spôsobmi. Prvý spôsob využíva jednostranné testy uvedené vyššie. Vyberieme menšiu z P-hodnôt a tú prenásobíme dvoma. Ak by však obe P-hodnoty boli väčšie ako $\frac{1}{2}$, P-hodnota dvojstranného testu by bola väčšia ako 1, preto v nasledovnom výraze zavádzame konštantu $\frac{1}{2}$: P-hodnota bude mať tvar $P = 2\min\{1/2, P_D, P_H\}$. V tomto prípade bude mať dvojstranný test dvojnásobne veľkú P-hodnotu ako jednostranné testy.

Pri použití druhého spôsobu výpočtu získame P-hodnotu ako súčet pravdepodobností rozloženia všetkých genotypov, ktorých pravdepodobnosť je menšia alebo rovná pravdepodobnosti skutočného rozloženia. Označme možný počet heterozygotov za podmienky, že x_A a x_a sú pevné ako j, cez ktoré budeme sčitovať. Potom j musí splňovať:

$$P(X_{AA} = x_{AA}, X_{Aa} = x_{Aa}, X_{aa} = x_{aa} \mid X_A = x_A)$$
$$\geq P(X_{AA} = \frac{x_A - j}{2}, X_{Aa} = j, X_{aa} = \frac{x_a - j}{2} \mid X_A = x_A)$$

Označme

$$\begin{aligned} N := \{j; x_A = 2x_{AA} + j, x_a = 2x_{aa} + j, x_{AA} + j + x_{aa} = n \\ \wedge P(X_{AA} = x_{AA}, X_{Aa} = x_{Aa}, X_{aa} = x_{aa} \mid X_A = x_A) \\ \geq P(X_{AA} = x_{AA}, X_{Aa} = j, X_{aa} = x_{aa} \mid X_A = x_A)\} \end{aligned} .$$

Potom P-hodnotu získame ako súčet:

$$P_\alpha = \sum_{j \in N} P(X_{AA} = x_{AA}, X_{Aa} = j, X_{aa} = x_{aa} \mid X_A = x_A).$$

3. Testy χ^2 pri neznámych parametroch

Uvažujme náhodný vektor $(X_1, ..., X_k)'$ s multinomickým rozdelením s pravdepodobnosťami $(p_1, ..., p_k)'$. Predpokladajme, že pravdepodobnosti p_1,...,p_k závisia na nejakom parametre $\theta = (\theta_1, ..., \theta_m)$, $p_1(\theta) + ... + p_k(\theta) = 1$, $0 < p_i < 1$, $i = 1, ...k$.
Odhadnime parameter θ metódou maximálnej vierohodnosti:
Definujme

$$L(\theta) := \ln\left(\frac{n!}{x_1!...x_k!} p_1(\theta)^{x_1} ... p_k(\theta)^{x_k}\right)$$

Potom platí:

$$L(\theta) = \ln\left(\frac{n!}{x_1!...x_k!}\right) + \sum_{i=1}^{k} x_i \ln p_i(\theta).$$

Z reťazového pravidla vyplýva:

$$\frac{\partial L(\theta)}{\partial \theta_j} = \sum_{i=1}^{k} \frac{x_i}{p_i(\theta)} \frac{\partial p_i(\theta)}{\partial \theta_j}$$

Riešením rovníc

$$\sum_{i=1}^{k} \frac{x_i}{p_i(\theta)} \frac{\partial p_i(\theta)}{\partial \theta_j} = 0 \tag{3.1}$$

pre $j = 1, ..., m$ dostávame odhad parametru θ metódou maximálnej vierohodnosti.
Ďalej označme

$$\chi^2(\theta) := \sum_{i=1}^{k} \frac{[X_i - np_i(\theta)]^2}{np_i(\theta)}$$

Veta 3.1. *Buď $m < k - 1$ a nech pre všetky θ z nedegenerovaného intervalu $\Theta \subset \mathbb{R}_m$ a pre všetky p_i platia nasledujúce predpoklady:*
(1) $p_1(\theta) + ... + p_k(\theta) = 1$.
(2) Existuje také $c > 0$, že pre každé $i = 1, ..., k$ platí $p_i(\theta) > c^2$.
(3) Pre každé $i = 1, ..., k$ má funkcia $p_i(\theta)$ spojité derivácie $\frac{\partial p_i(\theta)}{\partial \theta_s}$ a $\frac{\partial^2 p_i(\theta)}{\partial \theta_s \partial \theta_t}$ pre $s, t = 1, ..., m$
(4) Pre $M = \left(\frac{\partial p_i(\theta)}{\partial \theta_j}\right)_{k \times m}$ platí, že $h(M) = m$.
Nech sú možné výsledky náhodného pokusu Υ rozdelené do k nezlučitelných skupín a predpokladajme, že pravdepodobnosť toho, že výsledok prislúcha skupine i je $p_i^0 = p_i(\theta_1^0, ..., \theta_m^0)$, kde $\theta^0 = (\theta_1^0, ..., \theta_m^0)$ je vnútorným bodom intervalu Θ. Nech x_i určuje počet výsledkov prislúchajúcich k i-tej skupine, tj. po n opakovaniach náhodného pokusu Υ platí $\sum_{i=1}^{n} x_i = n$.
Potom rovnice (2.1) majú práve jeden systém riešení $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_n)$ taký, že $\tilde{\theta}$ konverguje k θ^0 pre $n \to \infty$. Po dosadení $\tilde{\theta}$ má veličina $\chi^2(\tilde{\theta}_n)$ asymptoticky χ^2_{k-m-1} rozdelenie pre $n \to \infty$.

Dôkaz. Cramér (1946). □

Majme populáciu s genotypmi AA, Aa, aa, náhodné veličiny X_{AA}, X_{Aa}, X_{aa} a pravdepodobnosti p_{AA}, p_{Aa}, p_{aa} definované v kapitole 1. Predpokladajme, že populácia je v Hardyho-Weinbergovej rovnováhe. V tejto situácií sú pravdepodobnosti p_{AA}, p_{Aa}, p_{aa} závislé na nejakom parametre θ a to spôsobom, že $p_{AA} = \theta^2$, $p_{Aa} = 2\theta(1-\theta)$, $p_{aa} = (1-\theta)^2$. Z predpokladov 3.1 plynie, že $m = 1$, t.j. θ je jednorozmerné.
V tomto prípade má $\chi^2(\theta)$ tvar:

$$\begin{aligned}\chi^2(\theta) &= \frac{[X_{AA} - np_{AA}(\theta)]^2}{np_{AA}(\theta)} + \frac{[X_{Aa} - np_{Aa}(\theta)]^2}{np_{Aa}(\theta)} + \frac{[X_{aa} - np_{aa}(\theta)]^2}{np_{aa}(\theta)} \\ &= \frac{[X_{AA} - n\theta^2]}{n\theta^2} + \frac{[X_{Aa} - 2n\theta(1-\theta)]}{2n\theta(1-\theta)} + \frac{[X_{aa} - n(1-\theta)^2]}{n(1-\theta)^2}\end{aligned} \tag{3.2}$$

Pri odhadovaní θ metódou maximálnej vierohodnosti dostávame rovnicu:

$$\frac{X_{AA}}{p_{AA}(\theta)}p'_{AA}(\theta) + \frac{X_{Aa}}{p_{Aa}(\theta)}p'_{Aa}(\theta) + \frac{X_{aa}}{p_{aa}(\theta)}p'_{aa}(\theta) = 0, \tag{3.3}$$

tj.

$$2\frac{X_{AA}}{\theta^2}\theta + \frac{X_{Aa}}{\theta(1-\theta)}(1-2\theta) - 2\frac{X_{aa}}{(1-\theta)^2}(1-\theta) = 0$$

Z toho dostávame odhad $\tilde{\theta_n} = \frac{X_A}{2n}$. Overme predpoklady vety 3.1:
Predpoklad (1): platí už z definovania problému.
Predpoklad (2): Z definície multinomického rozdelenia vieme, že $0 < p_i < 1$ pre $i \in \{AA, Aa, aa\}$. Z toho plynie, že $0 < \theta < 1$ a predpoklad (2) je splnený.
Predpoklad (3): Prvé a druhé derivácie majú tvar:

$$p'_{AA}(\theta) = 2\theta, \qquad p'_{Aa}(\theta) = 2(1-2\theta), \qquad p'_{aa}(\theta) = -2(1-\theta)$$

$$p''_{AA} = 2, \qquad p''_{Aa} = -4, \qquad p''_{aa} = 2$$

Prvé derivácie sú teda lineárne funkcie spojité na $(0,1)$ a druhé derivácie sú konštantné, teda aj spojité na celom intervale $(0,1)$.
Predpoklad (4): Vieme, že $m = 1$, tj. $M = (p'_{AA}(\theta), p'_{Aa}(\theta), p'_{aa}(\theta)) = (2\theta, 2(1-2\theta), -2(1-\theta))$ a M je teda nenulový vektor. Z toho vyplýva $h(M) = 1 = m$.
Z 3.1 plynie, že $\chi^2(\tilde{\theta_n})$ z (3.2), kde $\tilde{\theta_n}$ je riešenie (3.3), má asymptoticky χ^2_1 rozdelenie pri $n \to \infty$.
V praxi sa často využíva χ^2 test s opravou na spojitosť, z dôvodu, že multinomické, teda diskrétne rozdelenie aproximujeme rozdelením χ^2_1, čo je spojité rozdelenie. Rozdelenie χ^2 s opravou na spojitosť bude mať tvar:

$$\chi^2 = \frac{(\mid X_{AA} - np_{AA} \mid -c)^2}{np_{AA}} + \frac{(\mid X_{Aa} - np_{Aa} \mid -c)^2}{np_{Aa}} + \frac{(\mid X_{aa} - np_{aa} \mid -c)^2}{np_{aa}}$$

kde za c volíme najčastejšie hodnoty 0,5 alebo 0,25.

4. Modifikovaný χ^2 test

Uvažujme populáciu v Hardyho-Weinbergovej rovnováhe ako v kapitole 3. Predpokladajme, že x_A, x_a sú pevné. V kapitole 2 sme odvodili, že pravdepodobnosť javu $X_{AA} = x_{AA}, X_{Aa} = x_{Aa}, X_{aa} = x_{aa}$ za podmienky $X_A = x_A, X_a = x_a$ je:

$$P(X_{AA} = x_{AA}, X_{Aa} = x_{Aa}, X_{aa} = x_{aa} \mid X_A = x_A, X_a = x_a) = \frac{n! x_A! x a!}{2n! x_{AA}! x_{Aa}! x_{aa}!} 2^{x_{Aa}}$$

Pomocou tohto vzorca sme schopní vypočítať podmienené stredné hodnoty $E[X_{AA} \mid X_A = x_A, X_a = x_a]$, $E[X_{Aa} \mid X_A = x_A, X_a = x_a]$,
$E[X_{aa} \mid X_A = x_A, X_a = x_a]$.
Počítajme:

$$\begin{aligned} E[X_{AA} \mid X_A = x_A, X_a = x_a] &= \sum_{x_{AA}, x_{Aa}, x_{aa}} \frac{n! x_A! x_a!}{2n! x_{AA}! x_{Aa}! x_{aa}!} x_{AA} 2^{x_{Aa}} \\ &= \frac{n x_A (x_A - 1)}{2n(2n-1)} \underbrace{\sum_{x_{AA}, (x_{Aa}-1), x_{aa}} \frac{(n-1)!(x_A - 2)! x_a!}{(2n-2)!(x_{AA}-1)! x_{Aa}! x_{aa}!} 2^{x_{Aa}}}_{\star} \\ &= \frac{x_A (x_A - 1)}{2(2n-1)} \end{aligned}$$

$$\begin{aligned} E[X_{Aa} \mid X_A = x_A, X_a = x_a] &= \sum_{x_{AA}, x_{Aa}, x_{aa}} \frac{n! x_A! x_a!}{2n! x_{AA}! x_{Aa}! x_{aa}!} x_{Aa} 2^{x_{Aa}} \\ &= 2\frac{n x_A x_a}{2n(2n-1)} \underbrace{\sum_{x_{AA}, (x_{Aa}-1), x_{aa}} \frac{(n-1)!(x_A - 1)!(x_a - 1)!}{(2n-2)! x_{AA} (x_{Aa}-1)! x_a!} 2^{x_{Aa}-1}}_{\star} \\ &= \frac{x_A x_a}{(2n-1)} \end{aligned}$$

Pričom v oboch prípadoch v prvej sume sčítavame cez všetky x_{AA}, x_{Aa}, x_{aa}, pre ktoré platí: $x_{AA} + x_{Aa} + x_{aa} = n$ a v druhej cez všetky x_{AA}, $x_{Aa} - 1$, x_{aa}, pre ktoré platí: $x_{AA} + x_{Aa} - 1 + x_{aa} = n - 1$, $x_A = 2x_{AA} + x_{Aa}$, $x_a = 2x_{aa} + x_{Aa}$. Suma $\star$ je rovná jednej, keďže sčítame cez všetky možné pravdepodobnosti trinomického rozdelenia.
Rovnako ako pre $E[X_{AA} \mid X_A = x_A, X_a = x_a]$ by sme odvodili: $E[X_{aa} \mid X_A = x_A, X_a = x_a] = \frac{x_a(x_a-1)}{2(2n-1)}$.
Pomocou podmienených stredných hodnôt vyjadríme nový odhad θ^* parametru θ, pričom vyžadujeme aby $(\theta^*)^2 = nE[X_{AA} \mid X_A = x_A, X_a = x_a]$. Preto má θ^* tvar $\theta^* = \sqrt{\frac{X_A(X_A-1)}{2n(2n-1)}}$.
Dokážme, že:

$$\frac{X_A}{2n} \xrightarrow[n\to\infty]{P} \theta$$
$$\frac{X_A - 1}{2n - 1} \xrightarrow[n\to\infty]{P} \theta$$

Stačí ak dokážeme, že $\frac{X_A}{2n} \xrightarrow[n\to\infty]{L_2} \theta$(resp. $\frac{X_A-1}{2n-1} \xrightarrow[n\to\infty]{L_2} \theta$).

$$E\left|\theta - \frac{X_A}{2n}\right|^2$$
$$= E\left(\theta^2 - \frac{\theta X_A}{n} + \frac{X_A^2}{(2n)^2}\right) \xrightarrow{n\to\infty} 0$$
$$E\left|\theta - \frac{X_A - 1}{2n-1}\right|^2 \xrightarrow{n\to\infty} 0$$

Z tohto vyplýva, že $\tilde{\theta} \xrightarrow[n\to\infty]{P} \theta$ a $\theta^* \xrightarrow[n\to\infty]{P} \theta$. Teda $\tilde{\theta}$ a θ^* sú asymptoticky ekvivalentné.

Poznámka 1. *Veta 1 platí pre všetky asymtoticky normálne a asymptoticky efektívne odhady parametrov.*
[3, str.506]

Z Poznámky 1 plynie, že modifikovaný χ^2 test má pre $n \to \infty$ asymptoticky χ_1^2 rozdelenie.
Podobne ako v χ^2 teste aj tu môžeme použiť opravu na spojitosť. Potom modifikovaný χ^2 test s opravou na spojitosť bude mať tvar:

$$\frac{(\mid X_{AA} - np_{AA}(\theta^*) \mid -c)^2}{np_{AA}(\theta^*)} + \frac{(\mid X_{Aa} - np_{Aa}(\theta^*) \mid -c)^2}{np_{Aa}(\theta^*)} + \frac{(\mid X_{aa} - np_{aa}(\theta^*) \mid -c)^2}{np_{aa}(\theta^*)},$$

kde c je zvyčajne 0,5 alebo 0,25.

5. Test pomerom vierohodnosti

V predchádzajúcich kapitolách sme odhadli parameter θ metódou maximálnej vierohodnosti. Teraz túto metódu využime k odhadu pravdepodobností p_{AA}, p_{Aa}, p_{aa} vo všeobecnom prípade:

$$L^*(p_{AA}, p_{Aa}, p_{aa}) := \ln\left(\frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}p_{AA}^{x_{AA}}p_{Aa}^{x_{Aa}}p_{aa}^{x_{aa}}\right)$$

Využime metódu Lagrangeových multiplikátorov s podmienkou $p_{AA}+p_{Aa}+p_{aa}-1=0$. Definujme $G(p_{AA}, p_{Aa}, p_{aa}, \lambda) := L^*(p_{AA}, p_{Aa}, p_{aa}) - \lambda(p_{AA}+p_{Aa}+p_{aa}-1)$. Máme rovnice:

$$\begin{aligned}
\frac{\partial G}{\partial p_{AA}} &= \frac{x_{AA}}{p_{AA}} - \lambda = 0\\
\frac{\partial G}{\partial p_{Aa}} &= \frac{x_{Aa}}{p_{Aa}} - \lambda = 0\\
\frac{\partial G}{\partial p_{aa}} &= \frac{x_{aa}}{p_{aa}} - \lambda = 0\\
\frac{\partial G}{\partial \lambda} &= p_{AA} + p_{Aa} + p_{aa} - 1 = 0
\end{aligned}$$

Riešením dostávame maximálne vierohodné odhady pravdepodobností p_{AA}, p_{Aa}, p_{aa}:

$$\hat{p}_{AA} = \frac{x_{AA}}{n}, \qquad \hat{p}_{Aa} = \frac{x_{Aa}}{n}, \qquad \hat{p}_{aa} = \frac{x_{aa}}{n}, \qquad \lambda = n \tag{5.1}$$

Veta 5.1. *(Zehnaova– Princíp invariancie pre maximálne vierohodné odhady) Ak $\tilde{\theta}$ je maximálne vierohodný odhad parametru θ, potom $u(\tilde{\theta})$ je maximálne vierohodným odhadom parametrickej funkcie $u(\theta)$.*

Dôkaz. Anděl(2011). □

Zo Zehnaovej vety vyplýva, že v nasledujúcej vete môžeme použiť ľubovolnú, vzájomne jednoznačnú reparametrizáciu p_{AA}, p_{Aa}, p_{aa}. V tomto prípade je výhodné použiť reparametrizáciu:

$$p_{AA}(f,\theta) = \theta^2 + f(1-\theta)\theta \tag{5.2}$$
$$p_{Aa}(f,\theta) = 2(1-f)\theta(1-\theta) \tag{5.3}$$
$$p_{aa}(f,\theta) = (1-\theta)^2 + f\theta(1-\theta), \tag{5.4}$$

kde $f \in (-1,1)$, $\theta \in (0,1) \cap (\frac{-f}{1-f}, \frac{1}{1-f})$. Testujeme hypotézu H_0: $f = 0$ voči alternatívnej hypotéze H_1: $f \neq 0$. Za platnosti nulovej hypotézy odhad parametru θ je $\tilde{\theta} = \frac{x_A}{2n}$ a pre odhady funkcií $p_{AA}(f,\theta)$, $p_{Aa}(f,\theta)$,$p_{aa}(f,\theta)$ platí:

$$\begin{aligned}
\tilde{p}_{AA}(f,\theta) &= p_{AA}(0,\tilde{\theta}) = \left(\frac{x_A}{2n}\right)^2\\
\tilde{p}_{Aa}(f,\theta) &= p_{Aa}(0,\tilde{\theta}) = \frac{x_A x_a}{2n^2}\\
\tilde{p}_{aa}(f,\theta) &= p_{aa}(0,\tilde{\theta}) = \left(\frac{x_A}{2n}\right)^2
\end{aligned}$$

Vo všeobecnom prípade majú maximálne vierohodné odhady $\hat{p}_{AA}(f,\theta) = p_{AA}(\hat{f},\hat{\theta})$, $\hat{p}_{Aa}(f,\theta) = p_{Aa}(\hat{f},\hat{\theta})$, $\hat{p}_{aa}(f,\theta) = p_{aa}(\hat{f},\hat{\theta})$ funkcií $p_{AA}(f,\theta)$, $p_{Aa}(f,\theta)$, $p_{aa}(f,\theta)$ tvar ako v (5.1).

Veta 5.2. *Predpokladajme, že $\{f(x,\alpha), \alpha \in \Omega\}$ je regulárny systém hustôt s Fisherovou mierou informácie, α_0 je skutočná hodnota α a nech platia nasledujúce predpoklady:*
(i) Nech $\Omega \subset \mathbb{R}_m$, $m \geq 2$ ktorý obsahuje taký otvorený interval ω, že skutočná hodnota parametru α_0 patrí do ω.
(ii) Nech $\boldsymbol{X} = (X_1, ..., X_l)'$, kde X_i sú nezávislé, rovnako rozdelené náhodné veličiny s hustotou $f(x,\alpha)$ vzhladom k nejakej σ-konečnej miere μ.
(iii) Nech $M = \{x : f(x,\alpha) > 0\}$ nezávisí na α.
(iv) Nech pre $\alpha_1, \alpha_2 \in \Omega$. Potom $f(x,\alpha_1) = f(x,\alpha_2)[\mu] \Leftrightarrow \alpha_1 = \alpha_2$.
(v) Existuje derivácia: $\frac{\partial^3 f(x,\alpha)}{\partial\alpha_i\partial\alpha_j\partial\alpha_k}$, pre skoro všetky $x \in M$, pre všetky $\alpha \in \Omega$ a pre všetky $i,j,k = 1,...,m$.
(vi) Pre všetky $\alpha \in \Omega$ platí $\int_M f''_{ij}(x,\alpha)d\mu(x) = 0$, $i,j = 1,...,m$.
(vii) Pre všetky $i,j,k = 1,...,m$ existujú funkcie $M_{ijk}(x) \geq 0$ tak, že $E_{\alpha_0}M_{ijk}(X) < \infty$ a

$$\left|\frac{\partial^3 \ln f(x,\alpha)}{\partial\alpha_1\partial\alpha_j\partial\alpha_k}\right| \leq M_{ijk}(x),$$

pre všetky $\alpha \in \omega$ a pre skoro všetky $x \in M$.
(viii) Nech matica $J(\alpha)$ je spojitá v bode $\alpha = \alpha_0$.
Nech $1 \leq k < m$. Označme $\tau = (\alpha_1, ..., \alpha_k)'$, $\psi = (\alpha_{k+1}, ..., \alpha_m)$. Predpokladajme, že testujeme hypotézu $H_0 : \tau = \tau_0$. Označme $\hat{\alpha}_l = (\hat{\tau}'_l, \hat{\psi}'_l)'$ maximálne vierohodný odhad parametru α, ktorý nie je závislý na žiadnych iných podmienkach, $\tilde{\alpha}_l = (\tau_0, \tilde{\psi}_l)$ maximálne vierohodný odhad parametru θ za podmienky, že platí nulová hypotéza a nech α_0 z bodu (viii) má tvar $\alpha_0 = (\tau'_0, \psi'_0)'$.
Potom:

$$LR^* = 2[L(\hat{\alpha}_l) - L(\tilde{\alpha}_l)] \xrightarrow{d} \chi^2_k.$$

Dôkaz. Anděl(2011). □

V našej situácii budú mať odhady $\hat{\alpha}_l$, $\tilde{\alpha}_l$ tvar:

$$\hat{\alpha}_l = \left(\hat{f}, \hat{\theta}\right)'$$
$$\tilde{\alpha}_l = \left(0, \frac{x_A}{2n}\right)'$$

Overme predpoklady vety 5.2
Predpoklad (i): Položme $\Omega = \{(f,\theta); f \in (-1,1), \theta \in (0,1) \cap (\frac{-f}{1-f}, \frac{1}{1-f})\}$. Zvolme $\varepsilon > 0, \delta > 0$ dosť malé, tak aby množina $\omega = (-\varepsilon,\varepsilon) \times (\theta' - \delta, \theta' + \delta)$ bola podmnožinou množiny Ω, kde θ' je pravá hodnota parametru θ v prípade, že $f = 0$ (tj. platí $\theta' = \sqrt{p_{AA}}$). Tým máme predpoklad (i) overený.
Predpoklad (ii): Predpoklad platí, keďže uskutočňujeme nezávislé náhodné výbery z trinomického rozdelenia $M(1, p_{AA}, p_{Aa}, p_{aa})$.
Predpoklad (iii): $\frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}p_{AA}(f,\theta)^{x_{AA}}p_{Aa}(f,\theta)^{x_{Aa}}p_{aa}(f,\theta)^{x_{aa}}$ nadobúda len kladné hodnoty pre všetky $f, \theta \in \omega$, tj. množina M je nezávislá na f, θ.
Predpoklad(iv): Označme $F(f,\theta) := \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!}(\theta^2 + f\theta(1-\theta))^{x_{AA}}(2(1-f)\theta(1-\theta))^{x_{Aa}}((1-\theta)^2 + f\theta(1-\theta))^{x_{aa}}$. Nech pre všetky trojice x_{AA}, x_{Aa}, x_{aa} platí

$$F(f_1, \theta_1) = F(f_2, \theta_2)$$

Potom vhodnou voľbou trojíc x_{AA}, x_{Aa}, x_{aa} získame:

$$\begin{aligned}
\theta_1^2 + f_1\theta_1(1-\theta_1) &= \theta_2^2 + f_2\theta_2(1-\theta_2) \\
2(1-f_1)(1-\theta_1)\theta_1 &= 2(1-f_2)(1-\theta_2)\theta_2 \\
(1-\theta_1)^2 + f_1\theta_1(1-\theta_1) &= (1-\theta_2)^2 + f_2\theta_2(1-\theta_2)
\end{aligned}$$

Z druhého výrazu vyjadríme:

$$f_1(1-\theta_1)\theta_1 - f_2(1-\theta_2)\theta_2 = \theta_1(1-\theta_1) - \theta_2(1-\theta_2)$$

Po dosadení do prvého výrazu vieme odvodiť, že $\theta_1 = \theta_2$ a z tohto $f_1 = f_2$. Druhá implikácia, ktorá tvrdí, že ak $\theta_1 = \theta_2$ a $f_1 = f_2$ potom $F(f_1, \theta_1) = Ff_2, \theta_2$ je zrejme pravdivá.
Predpoklad (v): Pre x_{AA}, x_{Aa}, x_{aa} dosť veľké (napr. $x_{AA} \geq 3$, $x_{Aa} \geq 3$, $x_{aa} \geq 3$) tretie parciálne derivácie vyjdú ako lineárne kombinácie súčinov mocnín nejakých funkcií f a θ. Tieto funkcie sú tvorené súčtami súčinov prvkov f, θ, $1-f$, $1-\theta$, prípadne ich mocninami (o tomto sa môžeme presvedčiť napríklad programom *Mathematica*). Takéto výrazy sú definované pre všetky $f, \theta \in \omega$.
Predpoklad (vi): Z vety o zámene sumy a derivácie vyplýva:

$$\begin{aligned}
\int_M F''_{ij}(x, f, \theta)d\mu(x) &= \sum_{x_{AA}, x_{Aa}, x_{aa}} F''_{ij}((x_{AA}, x_{Aa}, x_{aa}), f, \theta) \\
&= \left(\sum_{x_{AA}, x_{Aa}, x_{aa}} F_{ij}((x_{AA}, x_{Aa}, x_{aa}), f, \theta) \right)'' \\
&= 1'' = 0,
\end{aligned}$$

pre $i, j \in \{AA, Aa, aa\}$.
Predpoklad (vii): Vypočítajme parciálne derivácie funkcie $\ln F$ podľa f a θ:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln F &= \frac{x_{AA}}{\theta} + \frac{x_{AA}}{(\theta + f(1-\theta))}(1-f) + \frac{x_{Aa}}{\theta} + \frac{x_{Aa}}{(1-\theta)} \\
&\quad + \frac{x_{aa}}{(1-\theta)} + \frac{x_{aa}}{(1-\theta+f\theta)}(f-1) \\
\frac{\partial^2}{\partial\theta^2}\ln F &= -\frac{x_{AA}}{\theta^2} - \frac{x_{AA}}{(\theta + f(1-\theta))^2}(1-f)^2 - \frac{x_{Aa}}{\theta^2} - \frac{x_{Aa}}{(1-\theta)^2} \\
&\quad - \frac{x_{aa}}{(1-\theta)^2} - \frac{x_{aa}}{(1-\theta+f\theta)^2}(f-1)^2 \\
\frac{\partial^3}{\partial\theta^3}\ln F &= 2\frac{x_{AA}}{\theta^3} + 2\frac{x_{AA}}{(\theta + f(1-\theta))^3}(1-f)^3 + 2\frac{x_{Aa}}{\theta^3} - 2\frac{x_{Aa}}{(1-\theta)^3} \\
&\quad - 2\frac{x_{aa}}{(1-\theta)^3} + 2\frac{x_{aa}}{(1-\theta+f\theta)^3}(f-1)^3 \\
\frac{\partial}{\partial f}\ln F &= \frac{x_{AA}}{\theta + f(1-\theta)}(1-\theta) - \frac{x_{Aa}}{1-f} + \frac{x_{aa}}{1-\theta+f\theta}\theta \\
\frac{\partial^2}{\partial f^2}\ln F &= -\left(\frac{x_{AA}}{(\theta + f(1-\theta))^2}(1-\theta)^2 + \frac{x_{Aa}}{(1-f)^2} + \frac{x_{aa}}{(1-\theta+f\theta)^2}\theta^2 \right)
\end{aligned}$$

$$\frac{\partial^3}{\partial f^3}\ln F = 2\frac{x_{AA}}{(\theta+f(1-\theta))^3}(1-\theta)^3 - 2\frac{x_{Aa}}{(1-f)^3} + 2\frac{x_{aa}}{(1-\theta+f\theta)^3}\theta^3$$
$$\frac{\partial^2}{\partial f\partial\theta}\ln F = \frac{-x_{AA}}{(\theta+f(1-\theta))^2} + \frac{x_{aa}}{(1-\theta+f\theta)^2}$$
$$\frac{\partial^3}{\partial\theta\partial f^2}\ln F = 2\left(\frac{x_{AA}(1-\theta)}{(\theta+f(1-\theta))^3} - \frac{x_{aa}\theta}{(1-\theta+f\theta)^3}\right)$$
$$\frac{\partial^3}{\partial f\partial\theta^2}\ln F = 2\left(\frac{x_{AA}(1-f)}{(\theta+f(1-\theta))^3} + \frac{x_{aa}(1-f)}{(1-\theta+f\theta)^3}\right)$$

Teraz potrebujeme obmedziť tretie parciálne derivácie $\ln F$. Použijeme fakt, že p_{AA}, p_{Aa}, p_{aa} sú parametrizované spôsobom (5.2) – (5.4). Z definície multinomického rozdelenia (konkrétnejšie z faktu, že $p_{AA}>0$, $p_{Aa}>0$, $p_{aa}>0$) vyplýva, že $\theta+f(1-\theta)>0 \wedge 1-\theta+f\theta>0$. Zvoľme ľubovolne malé $\beta>0$, $\gamma>0$. Potom môžeme predpokladať, že $\theta+f(1-\theta)>\beta \wedge 1-\theta+f\theta>\gamma$.

$$\begin{aligned}\left|\frac{\partial^3}{\partial\theta^3}\ln F\right| &= 2\left|\frac{x_{AA}}{\theta^3}\right| + 2\left|\frac{x_{AA}}{(\theta+f(1-\theta))^3}(1-f)^3\right| + 2\left|\frac{x_{Aa}}{\theta^3}\right| + 2\left|\frac{x_{Aa}}{(1-\theta)^3}\right| \\ &\quad + 2\left|\frac{x_{aa}}{(1-\theta)^3}\right| + 2\left|\frac{x_{aa}}{(1-\theta+f\theta)^3}(f-1)^3\right| \\ &\le 2\frac{x_{AA}}{(\theta'-\delta)^3} + 16\frac{x_{AA}}{\beta^3} + 2\frac{x_{Aa}}{(\theta'-\delta)^3} + 2\frac{x_{Aa}}{(1-\theta'-\delta)^3} \\ &\quad + 2\frac{x_{aa}}{(1-\theta'-\delta)^3} + 16\frac{x_{aa}}{\gamma^3} =: M_1(x_{AA},x_{Aa},x_{aa})\end{aligned}$$
$$\begin{aligned}\left|\frac{\partial^3}{\partial f^3}\ln F\right| &= 2\left|\frac{x_{AA}}{(\theta+f(1-\theta))^3}(1-\theta)^3\right| + 2\left|\frac{x_{Aa}}{(1-f)^3}\right| + 2\left|\frac{x_{aa}}{(1-\theta+f\theta)^3}\theta^3\right| \\ &\le 2\frac{x_{AA}}{\beta^3} + 2\frac{x_{Aa}}{(1-\varepsilon)^3} + 2\frac{x_{aa}}{\gamma^3} =: M_2(x_{AA},x_{Aa},x_{aa})\end{aligned}$$
$$\begin{aligned}\left|\frac{\partial^3}{\partial\theta\partial f^2}\ln F\right| &= 2\left|\frac{x_{AA}(1-\theta)}{(\theta+f(1-\theta))^3}\right| + 2\left|\frac{x_{aa}\theta}{(1-\theta+f\theta)^3}\right| \\ &\le 2\frac{x_{AA}}{\beta^3} + 2\frac{x_{aa}}{\gamma^3} =: M_3(x_{AA},x_{Aa},x_{aa})\end{aligned}$$
$$\begin{aligned}\left|\frac{\partial^3}{\partial f\partial\theta^2}\ln F\right| &= 2\left|\frac{x_{AA}(1-f)}{(\theta+f(1-\theta))^3}\right| + 2\left|\frac{x_{aa}(1-f)}{(1-\theta+f\theta)^3}\right| \\ &\le 4\frac{x_{AA}}{\beta^3} + 4\frac{x_{aa}}{\gamma^3} =: M_4(x_{AA},x_{Aa},x_{aa})\end{aligned}$$

Zrejme platí, že $M_i(x)\ge 0$, $i=1,...,4$. Nakoniec musíme overiť, že $E_{\alpha_0}M_i(X)<\infty$:

$$\begin{aligned}E_{\alpha_0}M_1(X) &= \frac{2n\theta^2}{(\theta'-\delta)^3} + \frac{16n}{\beta^3}\theta^2 + \frac{4n\theta(1-\theta)}{(\theta'-\delta)^3} + \frac{4n\theta(1-\theta)}{(1-\theta'-\delta)^3} + \frac{2n(1-\theta)^2}{(1-\theta'-\delta)^3} \\ &\quad + \frac{16n}{\gamma^3}(1-\theta)^2 < \infty\end{aligned}$$
$$E_{\alpha_0}M_2(X) = \frac{2n\theta^2}{\beta^3} + \frac{2n\theta(1-\theta)}{(1-\varepsilon)^3} + \frac{2n(1-\theta)^2}{\gamma^3} < \infty$$
$$E_{\alpha_0}M_3(X) = \frac{2n\theta^2}{\beta^3} + \frac{2n(1-\theta)^2}{\gamma^3} < \infty$$

$$E_{\alpha_0} M_4(X) = \frac{4n\theta^2}{\beta^3} + \frac{4n(1-\theta)^2}{\gamma^3} < \infty$$

Predpoklad (viii): Pomocou výsledkov z predchádzajúceho bodu môžeme vypočítať Fisherovu informačnú maticu.

$$\begin{aligned}
J_{11} &= -\sum_{x_{AA},x_{Aa},x_{aa}} \frac{\partial^2 \ln F}{\partial\theta^2} F = \sum_{x_{AA},x_{Aa},x_{aa}} F\big(\frac{x_{AA}}{\theta^2} + \frac{x_{AA}}{(\theta + f(1-\theta))^2}(1-f)^2 + \frac{x_{Aa}}{\theta^2} \\
&+ \frac{x_{Aa}}{(1-\theta)^2} + \frac{x_{aa}}{(1-\theta)^2} + \frac{x_{aa}}{(1-\theta+f\theta)^2}(f-1)^2\big) \\
&= EX_{AA}\left(\frac{1}{\theta^2} + \frac{(1-f)^2}{(\theta+f(1-\theta))^2}\right) + EX_{Aa}\left(\frac{1}{\theta^2} + \frac{1}{(1-\theta)^2}\right) \\
&+ EX_{aa}\left(\frac{1}{(1-\theta)^2} + \frac{(f-1)^2}{(1-\theta+f\theta)^2}\right) \\
&= \frac{n(\theta+f(1-\theta))}{\theta} + \frac{n\theta(1-f)^(2)}{\theta+f(1-\theta)} + \frac{2n(1-f)(1-\theta)}{\theta} + \frac{2n(1-f)\theta}{1-\theta} \\
&+ \frac{n(1-\theta+f\theta)}{1-\theta} + \frac{n(1-\theta)}{1-\theta+f\theta}(f-1)^2
\end{aligned}$$

Pričom výsledok získame využitím vzorcov získaných z kapitoly 1. Podobne odvodíme ostatné prvky:

$$\begin{aligned}
J_{21} = J_{12} &= -\sum_{x_{AA},x_{Aa},x_{aa}} \frac{\partial^2 \ln F}{\partial f \partial\theta} F \\
&= \sum_{x_{AA},x_{Aa},x_{aa}} F\left(x_{AA}\frac{1}{(\theta+f(1-\theta))^2} - x_{aa}\frac{1}{(1-\theta+f\theta)^2}\right) \\
&= \frac{n\theta}{\theta+f(1-\theta)} - \frac{n(1-\theta)}{1-\theta+f\theta} \\
J_{22} &= -\sum_{x_{AA},x_{Aa},x_{aa}} \frac{\partial^2 \ln F}{\partial f^2} F \\
&= \sum_{x_{AA},x_{Aa},x_{aa}} F\left(\frac{x_{AA}}{(\theta+f(1-\theta))^2}(1-\theta)^2 + \frac{x_{Aa}}{(1-f)^2} + \frac{x_{aa}}{(1-\theta+f\theta)^2}\theta^2\right) \\
&= \frac{n\theta(1-\theta)^2}{\theta+f(1-\theta)} + \frac{2n\theta(1-\theta)}{1-f} + \frac{n(1-\theta)\theta^2}{1-\theta+f\theta}
\end{aligned}$$

Preto bude Fisherová informačná matica v bode $(0,\theta')$ spojitá a bude mať nasledovný tvar:

$$\begin{pmatrix} \frac{2n}{\theta'(1-\theta')} & 0 \\ 0 & n \end{pmatrix}$$

Tým sme overili predpoklady Vety 5.2. Preto

$$\begin{aligned}
&LR^* = 2[L(\hat{\theta}_n) - L(\tilde{\theta}_n)] = x_{AA}\ln x_{AA} + x_{Aa}\ln x_{Aa} + x_{aa}\ln x_{aa} \\
&- x_A \ln x_A - x_a \ln x_a - n\ln n + 2n\ln 2n \xrightarrow{n\to\infty} \chi^2_1
\end{aligned}$$

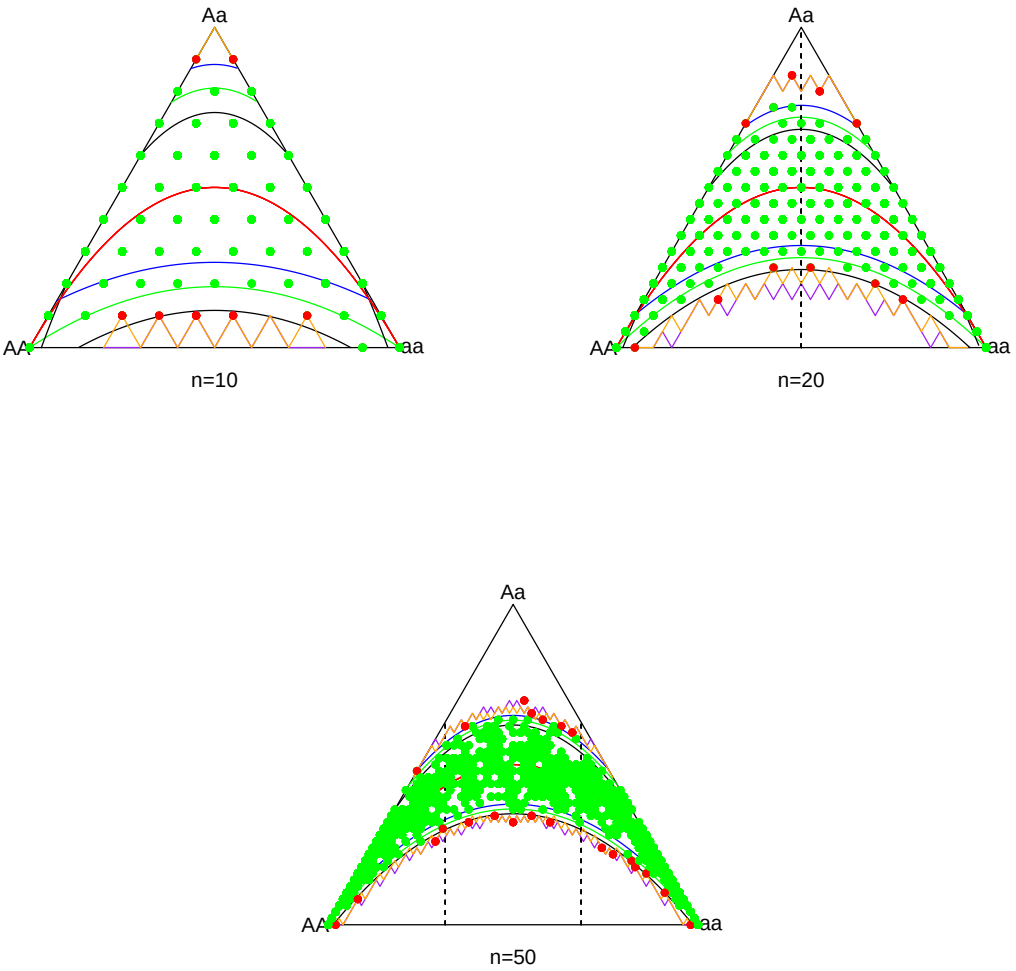
6. Porovnávanie testov

V nasledujúcej kapitole budeme používať program R a knižnicu HardyWeinberg (zdrojové kódy sa nachádzajú v prílohe č.1 – prílohe č. 5). Nami používané funkcie sú HWData, HWExact, HWTernaryplot. HWData používame pri simulácii dát, HWExact, HWTernaryplot pri testovaní nulových hypotéz. HWTernaryplot slúži na grafické znázornenie niektorých testov.

6.1 Grafické znázornenie niektorých testov

De Finettiho diagram

Obr. 6.1: De Finettiho diagramy pre $n = 10$, $n = 20$, $n = 50$



Majme rovnostranný trojholník XYZ a nech výška trojholníka je 1. Nech (p_{AA}, p_{Aa}, p_{aa}) sú pravdepodobnosti výskytu genotypov AA, Aa, aa. Po-

tom populáciu môžeme reprezentovať ako bod P v trojholníku XYZ, kde priesečnice kolmíc z bodu P na jednotlivé strany majú vzdialenosť od bodu P p_{AA}, p_{Aa}, p_{aa}. Predpokladajme, že bod Y reprezentuje heterozygotov, bod X homozygotov AA a bod Z homozygotov aa. Potom dĺžka kolmice, ktorá spája bod P so stranou XZ je p_{Aa}, dĺžka kolmice, ktorá spája bod P so stranou XY je p_{aa} a dĺžka kolmice, ktorá spája bod P so stranou YZ je p_{AA}. V prípade, že populácia je v Hardyho-Weinbergovej rovnováhe platí $p_{Aa}^2 = 4p_{AA}p_{aa}$ a body znázorňujúce pravdepodobnosti p_{AA}, p_{Aa}, p_{aa} ležia na parabole. Takýto diagram nazývame de Finettiho diagram.

Pomocou programu R, funkcie HWTernaryPlot, sme schopní zakresliť do de Finettiho diagramu veľký počet dát. Ak chceme graficky testovať hypotézu H_0 máme možnosť do diagramu zakresliť aj oblasti prijímania hypotézy H_0, ktorú môžeme určiť pomocou χ^2 testu, χ^2 testu s opravou na spojitosť a exaktného testu s oboma spôsobmi výpočtu P-hodnoty. Simulujme 1000 možných rozložení postupne pre $n = 10$, $n = 20$, $n = 50$ ktoré sú v Hardyho-Weinbergovej rovnováhe pomocou funkcie HWData. De Finettiho diagramy, pre rôzne n sú znázornené na obrázku 6.1. Zelené body znázorňujú populácie, pre ktoré nulovú hypotézu nezamietame, červené znázorňujú populácie, pre ktoré nulovú hypotézu zamietame. Krajné paraboly ohraničujú oblasť prijímania nulovej hypotézy. Označme $D := (x_{Aa} - EX_{Aa})/2$, kde EX_{Aa} sú stredné hodnoty heterozygotov v prípade, že populácia je v Hardyho-Weinbergovej rovnováhe. Potom krivka znázorňujúca trojice pravdepodobností, ktoré sú v Hardyho-Weinbergovej rovnováhe má červenú farbu, krivky, ktoré ohraničujú oblasť prijímania nulovej hypotézy pre χ^2 test majú zelenú farbu, pre χ^2 test s opravou na spojitosť $c = 0{,}5$ majú modrú farbu ak $D > 0$ alebo čiernu ak $D < 0$ a krivky, ktoré ohraničujú oblasť prijímania pre exaktný test s prvým spôsobom výpočtu P-hodnoty majú fialovú, s druhým spôsobom výpočtu P-hodnoty oranžovú farbu.

6.2 Hladina testov

V nasledujúcej časti sa budeme venovať odhadovaniu hladiny testov. Ako prvý krok si musíme simulovať dáta. Budeme generovať rozloženie genotypov pri veľkosti populácie $n = 100$, ktoré splňujú Hardyho-Weinbergovú rovnováhu. Na tento účel použijeme funkciu HWData, pričom budeme požadovať 1000 možných simulácií. Postupne uvažujme populácie s $p \in \{0,0{,}1, ..., 0{,}9, 1\}$. Ďalšie argumenty funkcie budú *exactequilibrium=FALSE* a $f = 0$, pričom prvým argumentom zabezpečíme, že počty genotypov, ktoré získame z HWData budú celé čísla.
Na takto získanú tabuľku hodnôt postupne aplikujeme vyššie zmienené štatistické testy. V prípade exaktného testu sme využívali oba spôsoby výpočtu P-hodnoty. χ^2 testy s opravou na spojitosť prevádzame najprv pre $c = 0{,}5$, potom pre $c = 0{,}25$. Nech $\alpha = 0{,}05$. Pre každý test sme zistili, že koľkokrát sme nulovú hypotézu zamietli (tj. koľkokrát bolo $P < \alpha$), pomocou tohto sme odhadli pravdepodobnosť chyby I. druhu, viz tabuľka 6.2.

Pomocou týchto výsledkov môžeme určiť 95% interval spoľahlivosti pravdepodobností chyby I. druhu. Intervalové odhady budú mať tvar viz tabuľka 6.2. Pre lepšiu názornosť výsledkov sme odhady pravdepodobností I. druhu, ktorých interval spoľahlivosti nepokrýva hodnotu $\alpha = 0{,}05$ a odhad je menší než α sme

test $\backslash\theta$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Exaktný 1.		0.009	0.034	0.03	0.031	0.032	0.027	0.029	0.023	0.015	
Exaktný 2.		0.018	0.041	0.035	0.037	0.043	0.034	0.036	0.033	0.024	
χ^2 test		0.036	0.057	0.042	0.051	0.052	0.044	0.044	0.046	0.047	
χ^2 $c = 0.5$		0.011	0.035	0.031	0.033	0.034	0.031	0.033	0.023	0.018	
χ^2 $c = 0.25$		0.023	0.041	0.036	0.039	0.037	0.035	0.036	0.032	0.036	
Mod. χ^2		0.037	0.053	0.043	0.051	0.049	0.046	0.044	0.043	0.047	
Mod. χ^2 $c = 0.5$		0.012	0.03	0.03	0.036	0.033	0.032	0.029	0.028	0.02	
Mod. χ^2 $c = 0.25$		0.023	0.044	0.037	0.037	0.043	0.037	0.036	0.037	0.036	
T. pom. vierohod.		0.026	0.069	0.047	0.051	0.052	0.045	0.048	0.064	0.034	

Tabuľka 6.1: Odhad hladiny testov

test $\backslash\theta$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Exaktný 1.		(0.007, 0.011)	(0.031, 0.038)	(0.027, 0.034)	(0.028, 0.035)	(0.029, 0.036)	(0.024, 0.03)	(0.026, 0.032)	(0.02, 0.026)	(0.013, 0.018)	
Exaktný 2.		(0.015, 0.021)	(0.037, 0.045)	(0.031, 0.039)	(0.033, 0.041)	(0.039, 0.047)	(0.031, 0.038)	(0.032, 0.04)	(0.03, 0.037)	(0.021, 0.027)	
χ^2 test		(0.032, 0.04)	(0.053, 0.062)	(0.038, 0.046)	(0.047, 0.055)	(0.048, 0.057)	(0.04, 0.048)	(0.04, 0.048)	(0.042, 0.05)	(0.043, 0.051)	
χ^2 $c = 0.5$		(0.009, 0.013)	(0.031, 0.039)	(0.028, 0.035)	(0.03, 0.037)	(0.031, 0.038)	(0.028, 0.035)	(0.03, 0.037)	(0.02, 0.026)	(0.015, 0.021)	
χ^2 $c = 0.25$		(0.02, 0.026)	(0.037, 0.045)	(0.032, 0.04)	(0.035, 0.043)	(0.033, 0.041)	(0.031, 0.039)	(0.032, 0.04)	(0.029, 0.036)	(0.032, 0.04)	
Mod. χ^2		(0.033, 0.041)	(0.049, 0.058)	(0.039, 0.047)	(0.047, 0.055)	(0.045, 0.053)	(0.042, 0.05)	(0.04, 0.048)	(0.039, 0.047)	(0.043, 0.051)	
Mod. χ^2 $c = 0.5$		(0.01, 0.014)	(0.027, 0.034)	(0.027, 0.034)	(0.032, 0.04)	(0.03, 0.037)	(0.029, 0.036)	(0.026, 0.032)	(0.025, 0.031)	(0.017, 0.023)	
Mod. χ^2 $c = 0.25$		(0.02, 0.026)	(0.04, 0.048)	(0.033, 0.041)	(0.033, 0.041)	(0.039, 0.047)	(0.033, 0.041)	(0.032, 0.04)	(0.033, 0.041)	(0.032, 0.04)	
T. pom. vierohod.		(0.023, 0.029)	(0.064, 0.074)	(0.043, 0.051)	(0.047, 0.055)	(0.048, 0.057)	(0.041, 0.049)	(0.044, 0.052)	(0.059, 0.069)	(0.031, 0.038)	

Tabuľka 6.2: Intervaly spoľahlivosti pravdepodobností I. druhu

označili zelenou farbou a odhady ktorých interval spoľahlivosti nepokrýva hodnotu $\alpha = 0{,}05$ a odhad je väčší než α sme označili červenou farbou. Vidíme, že najkonzervatívnejšie sú: exaktný test s prvým spôsobom výpočtu P-hodnoty, χ^2 test s opravou na spojitosť $c = 0{,}5$ a modifikovaný χ^2 test s opravou na spojitosť $c = 0{,}5$. Najlepšie z hladiska hladiny testu si vedie modifikovaný χ^2 test. Test pomerom vierohodnosti sa javí v dvoch, χ^2 test v jednom prípade antikonzervatívny.

6.3 Sila testov

Ďalšou charachteristikou testov, ktorú budeme odhadovať, je ich sila. V tomto prípade bude postup nasledovný: simulujeme dáta, ktoré nesplňujú Hardyho-Weinbergovú rovnováhu, zistíme, že v koľkých prípadoch daný test nulovú hypotézu zamietol a pomocou tejto informácie odhadneme pravdepodobnosť zamietnutia nulovej hypotézy za predpokladu, že hypotéza by mala byť zamietnutá. Generujme dáta v závislosti na f a p. Nech $f \in \{-0{,}9, ..., -0{,}1, 0{,}1, ..., 0{,}9\}$ a $\theta \in \{0, 0{,}1, ..., 0{,}9, 1\}$, $n = 100$, počet simulácií bude 1000. Potom odhad sily testov v závislosti od f a θ bude nasledovná:

$\theta \backslash f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.152	0.397	0.635	0.855	0.937	0.974	0.997	1	1
0.2								0.451	0.075	0.115	0.395	0.697	0.925	0.992	0.999	1	1	1
0.3						0.999	0.869	0.438	0.107	0.116	0.437	0.81	0.963	0.998	1	1	1	1
0.4				1	1	0.983	0.81	0.444	0.117	0.123	0.438	0.798	0.974	0.997	1	1	1	1
0.5	1	1	1	1	1	0.983	0.851	0.465	0.135	0.132	0.461	0.804	0.979	0.999	1	1	1	1
0.6				1	0.999	0.973	0.814	0.429	0.118	0.128	0.435	0.801	0.98	1	1	1	1	1
0.7						0.999	0.87	0.442	0.1	0.137	0.426	0.772	0.959	0.996	1	1	1	1
0.8								0.449	0.076	0.123	0.389	0.706	0.92	0.987	0.998	1	1	1
0.9									0	0.073	0.303	0.557	0.774	0.906	0.974	0.987	0.999	1
1																		

Tabuľka 6.3: Odhad sily exaktného testu s 1. definíciou P-hodnoty

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.138	0.38	0.654	0.845	0.94	0.986	0.998	0.999	1
0.2								0.496	0.101	0.152	0.454	0.753	0.948	0.993	0.999	1	1	1
0.3						0.999	0.882	0.468	0.13	0.14	0.471	0.834	0.97	0.999	1	1	1	1
0.4				1	1	0.985	0.835	0.48	0.132	0.139	0.476	0.837	0.979	0.998	1	1	1	1
0.5	1	1	1	1	1	0.985	0.863	0.489	0.146	0.155	0.525	0.841	0.985	0.999	1	1	1	1
0.6				1	0.999	0.98	0.836	0.462	0.14	0.149	0.473	0.834	0.985	1	1	1	1	1
0.7						0.999	0.884	0.472	0.113	0.158	0.469	0.796	0.966	0.999	1	1	1	1
0.8								0.495	0.093	0.153	0.449	0.754	0.939	0.99	1	1	1	1
0.9									0	0.118	0.399	0.645	0.836	0.934	0.984	0.993	1	1
1																		

Tabuľka 6.4: Odhad sily exaktného testu s 2. definíciou P-hodnoty

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.22	0.467	0.723	0.897	0.967	0.989	0.999	0.999	1
0.2								0.597	0.151	0.191	0.509	0.787	0.966	0.995	0.999	1	1	1
0.3						0.999	0.915	0.545	0.162	0.152	0.497	0.85	0.975	0.999	1	1	1	1
0.4				1	1	0.99	0.873	0.558	0.177	0.151	0.505	0.853	0.983	0.998	1	1	1	1
0.5	1	1	1	1	1	0.99	0.899	0.543	0.193	0.158	0.527	0.843	0.986	0.999	1	1	1	1
0.6				1	0.999	0.986	0.886	0.536	0.177	0.164	0.491	0.853	0.989	1	1	1	1	1
0.7						0.999	0.923	0.559	0.167	0.174	0.494	0.818	0.973	0.999	1	1	1	1
0.8								0.606	0.151	0.18	0.498	0.792	0.952	0.994	1	1	1	1
0.9									0	0.171	0.482	0.722	0.895	0.951	0.996	0.995	1	1
1																		

Tabuľka 6.5: Odhad sily χ^2 testu

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.116	0.333	0.603	0.823	0.925	0.977	0.997	0.999	1
0.2								0.431	0.071	0.117	0.398	0.704	0.926	0.993	0.999	1	1	1
0.3						0.999	0.882	0.471	0.13	0.117	0.437	0.81	0.963	0.998	1	1	1	1
0.4				1	1	0.986	0.844	0.487	0.136	0.121	0.437	0.797	0.974	0.997	1	1	1	1
0.5	1	1	1	1	1	0.984	0.859	0.476	0.138	0.129	0.456	0.8	0.979	0.999	1	1	1	1
0.6				1	0.999	0.981	0.841	0.47	0.144	0.127	0.429	0.795	0.98	1	1	1	1	1
0.7						0.999	0.883	0.472	0.114	0.137	0.426	0.772	0.959	0.996	1	1	1	1
0.8								0.415	0.072	0.124	0.393	0.712	0.923	0.989	0.998	1	1	1
0.9									0	0.092	0.356	0.602	0.812	0.924	0.981	0.99	1	1
1																		

Tabuľka 6.6: Odhad sily χ^2 testu s opravou na spojitosť $c = 0{,}5$

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.167	0.417	0.685	0.863	0.957	0.987	0.998	0.999	1
0.2								0.499	0.105	0.151	0.455	0.753	0.948	0.993	0.999	1	1	1
0.3						0.999	0.904	0.506	0.146	0.136	0.472	0.837	0.972	0.999	1	1	1	1
0.4				1	1	0.986	0.855	0.508	0.15	0.133	0.462	0.823	0.977	0.997	1	1	1	1
0.5	1	1	1	1	1	0.987	0.874	0.506	0.157	0.134	0.468	0.807	0.98	0.999	1	1	1	1
0.6				1	0.999	0.983	0.859	0.493	0.159	0.146	0.463	0.823	0.984	1	1	1	1	1
0.7						0.999	0.904	0.521	0.127	0.159	0.471	0.802	0.968	0.999	1	1	1	1
0.8								0.5	0.099	0.153	0.446	0.755	0.939	0.99	1	1	1	1
0.9									0	0.141	0.428	0.68	0.859	0.941	0.989	0.993	1	1
1																		

Tabuľka 6.7: Odhad sily χ^2 testu s opravou na spojitosť $c = 0{,}25$

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.221	0.468	0.724	0.9	0.967	0.992	0.999	1	1
0.2								0.55	0.135	0.195	0.518	0.791	0.967	0.995	0.999	1	1	1
0.3						0.999	0.915	0.545	0.162	0.156	0.502	0.854	0.977	0.999	1	1	1	1
0.4				1	1	0.989	0.866	0.53	0.168	0.157	0.528	0.861	0.983	0.999	1	1	1	1
0.5	1	1	1	1	1	0.987	0.883	0.518	0.167	0.166	0.548	0.854	0.987	1	1	1	1	1
0.6				1	0.999	0.985	0.872	0.513	0.173	0.175	0.508	0.864	0.99	1	1	1	1	1
0.7						0.999	0.923	0.559	0.166	0.176	0.505	0.824	0.973	0.999	1	1	1	1
0.8								0.562	0.131	0.183	0.502	0.794	0.952	0.994	1	1	1	1
0.9									0	0.174	0.483	0.723	0.896	0.955	0.996	0.996	1	1
1																		

Tabuľka 6.8: Odhad sily modifikovaného χ^2 testu

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.129	0.361	0.627	0.834	0.929	0.978	0.998	0.999	1
0.2								0.359	0.051	0.143	0.442	0.735	0.941	0.993	0.999	1	1	1
0.3						0.999	0.869	0.437	0.106	0.119	0.445	0.814	0.967	0.998	1	1	1	1
0.4				1	1	0.984	0.821	0.455	0.121	0.131	0.462	0.823	0.977	0.997	1	1	1	1
0.5	1	1	1	1	1	0.983	0.852	0.469	0.136	0.132	0.468	0.807	0.98	0.999	1	1	1	1
0.6				1	0.999	0.976	0.824	0.441	0.124	0.146	0.463	0.823	0.984	1	1	1	1	1
0.7						0.999	0.87	0.442	0.102	0.139	0.442	0.779	0.96	0.998	1	1	1	1
0.8								0.351	0.058	0.148	0.435	0.741	0.936	0.99	0.998	1	1	1
0.9									0	0.106	0.376	0.628	0.825	0.931	0.982	0.991	1	1
1																		

Tabuľka 6.9: Odhad sily modifikovaného χ^2 testu s opravou na spojitosť $c = 0{,}5$

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0	0.167	0.417	0.685	0.863	0.959	0.989	0.999	0.999	1
0.2								0.473	0.095	0.165	0.469	0.766	0.954	0.993	0.999	1	1	1
0.3						0.999	0.9	0.492	0.14	0.144	0.492	0.846	0.973	0.999	1	1	1	1
0.4				1	1	0.986	0.844	0.487	0.136	0.147	0.493	0.844	0.981	0.998	1	1	1	1
0.5	1	1	1	1	1	0.984	0.859	0.476	0.139	0.146	0.512	0.83	0.984	0.999	1	1	1	1
0.6				1	0.999	0.981	0.841	0.47	0.144	0.156	0.483	0.841	0.985	1	1	1	1	1
0.7						0.999	0.897	0.492	0.12	0.168	0.487	0.811	0.97	0.999	1	1	1	1
0.8								0.477	0.092	0.163	0.463	0.765	0.941	0.99	1	1	1	1
0.9									0	0.141	0.428	0.68	0.86	0.942	0.989	0.994	1	1
1																		

Tabuľka 6.10: Odhad sily modifikovaného χ^2 testu s opravou na spojitosť $c = 0{,}25$

$\theta \setminus f$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0																		
0.1									0.079	0.14	0.38	0.654	0.845	0.94	0.986	0.998	0.999	1
0.2								0.714	0.225	0.156	0.456	0.753	0.948	0.993	0.999	1	1	1
0.3						1	0.933	0.581	0.189	0.152	0.494	0.847	0.975	0.999	1	1	1	1
0.4				1	1	0.99	0.874	0.559	0.177	0.151	0.505	0.853	0.983	0.998	1	1	1	1
0.5	1	1	1	1	1	0.99	0.899	0.543	0.193	0.158	0.527	0.843	0.986	0.999	1	1	1	1
0.6				1	0.999	0.986	0.886	0.537	0.177	0.164	0.491	0.853	0.989	1	1	1	1	1
0.7						1	0.945	0.607	0.201	0.175	0.49	0.815	0.972	0.999	1	1	1	1
0.8								0.725	0.233	0.155	0.45	0.756	0.939	0.99	1	1	1	1
0.9									0.061	0.121	0.399	0.645	0.836	0.934	0.984	0.993	1	1
1																		

Tabuľka 6.11: Odhad sily testu pomerom vierohodnosti

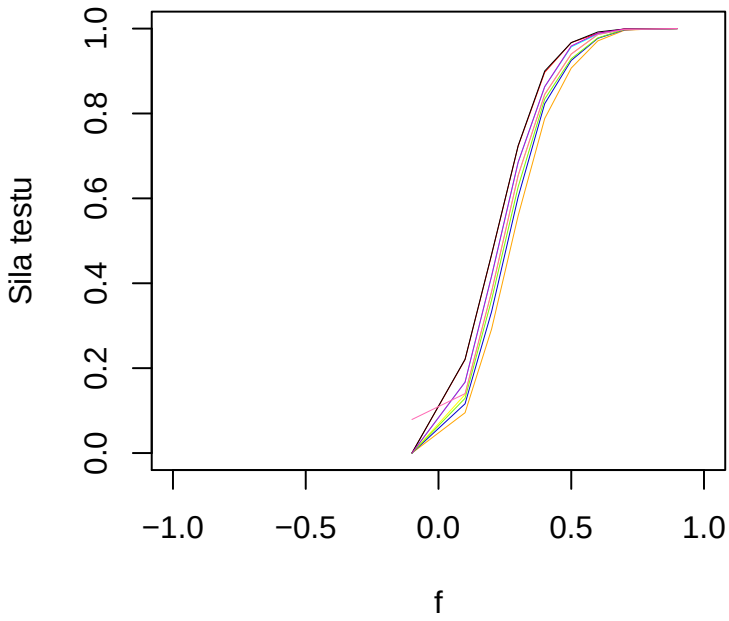
Prázdne miesta v tabuľkách značia, že takáto kombinácia parametrov f a θ nie je prípustná, tj. $(f, \theta) \notin \Omega$.

Porovnaním týchto tabuliek zistíme: χ^2 test je pre každú dvojicu (f, θ) silnejší alebo rovnako silný ako exaktný test, kde P-hodnotu počítame druhým spôsobom, ten je však silnejší alebo rovnako silný ako exaktný test, kde P-hodnotu počítame prvým spôsobom. Ďalej χ^2 test je pre každé uvedené (f, θ) silnejší alebo rovnako silný ako χ^2 s opravou na spojitosť $c = 0{,}25$, ktorý je silnejší ako χ^2 s

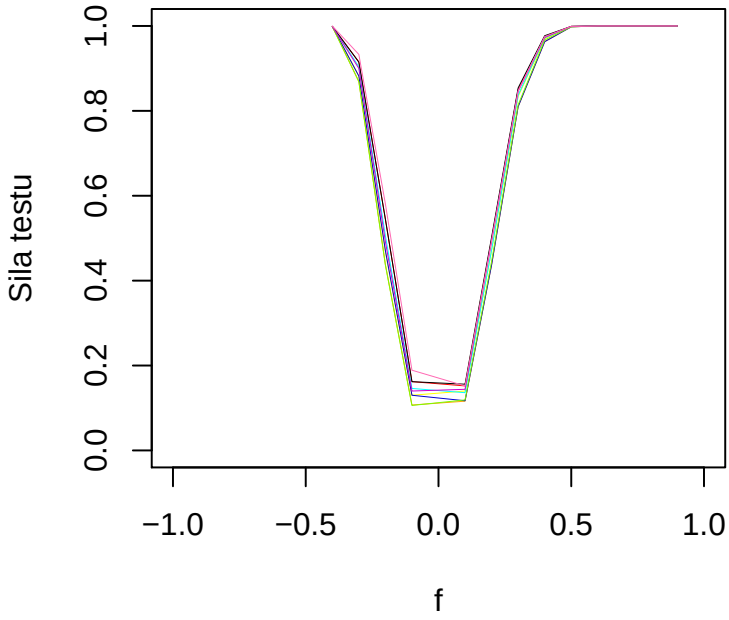
opravou na spojitosť $c = 0{,}5$. Taktiež platí, že χ^2 test je silnejší alebo rovnako silný ako modifikovaný χ^2 test s opravou na spojitosť $c = 0{,}25$, ktorý je silnejší ako modifikovaný χ^2 s opravou na spojitosť $c = 0{,}5$. Ďalej sme zistili, že pre všetky skupiny simulovaných dát platí, že modifikovaný χ^2 test s opravou na spojitosť $c = 0{,}25$ je silnejší alebo rovnako silný ako exaktný test s prvým spôsobom výpočtu P-hodnoty. Podobne ako pre χ^2 test, aj pre modifikovaný χ^2 test platí, že je silnejší alebo rovnako silný ako χ^2 s opravou na spojitosť $c = 0{,}25$, ktorý je silnejší ako χ^2 s opravou na spojitosť $c = 0{,}5$, je tiež silnejší alebo rovnako silný ako modifikovaný χ^2 s opravou na spojitosť $c = 0{,}25$, ktorý je silnejší ako modifikovaný χ^2 s opravou na spojitosť $c = 0{,}5$ a je silnejší ako oba exaktné testy. Test pomerom vierohodnosti je silnejší ako nasledovné testy: χ^2 s opravou na spojitosť $c = 0{,}5$, modifikovaný χ^2 s opravou na spojitosť $c = 0{,}5$ a oba exaktné testy. Zvyšné prípady nevieme jednoznačne porovnať.

Pre názornejšiu reprezentáciu výsledkov ich môžeme zobraziť graficky. Načrtnime grafy pre hodnoty $p = 0{,}1, 0{,}3, 0{,}5, 0{,}7, 0{,}9$. Oranžová farba prislúcha exaktnému testu s prvým spôsobom výpočtu P-hodnoty, žltá exaktnému testu s druhým spôsobom výpočtu P-hodnoty, červená χ^2 testu, tmavomodrá χ^2 testu s opravou na spojitosť $c = 0{,}5$, bledomodrá χ^2 testu s opravou na spojitosť $c = 0{,}25$, čierna modifikovanému χ^2 testu, zelená modifikovanému χ^2 testu s opravou na spojitosť $c = 0{,}5$, fialová modifikovanému χ^2 testu s opravou na spojitosť $c = 0{,}25$ a ružová testu pomerom vierohodnosti.

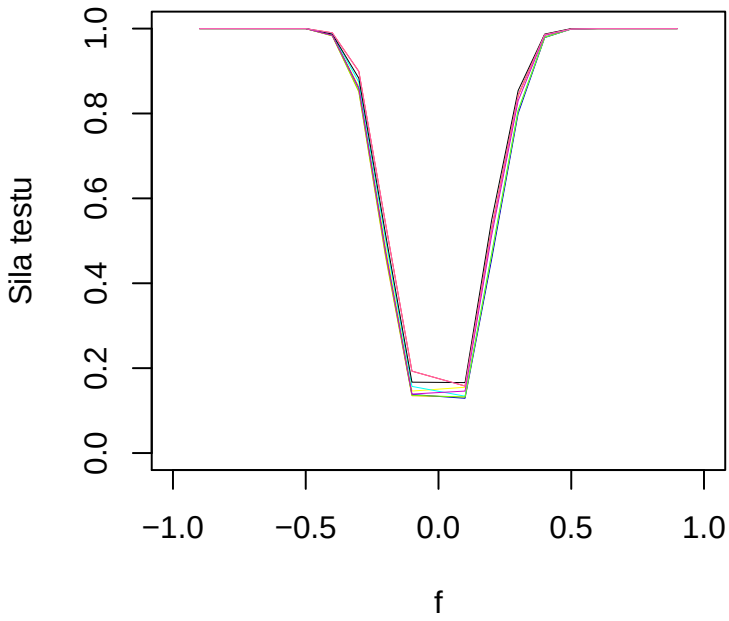
Zhrnutím časti o odhadovaní hladiny testov a o sile testov môžeme usudíť, že najlepšie výsledky pre $n = 100$ dosahuje modifikovaný χ^2 test. Naopak ako najmenej vyhovujúce pre $n = 100$ môžeme označiť exaktný test s prvým spôsobom výpočtu P-hodnoty, χ^2 test s opravou na spojitosť $c = 0{,}5$ a modifikovaný χ^2 test s opravou na spojitosť $c = 0{,}5$.



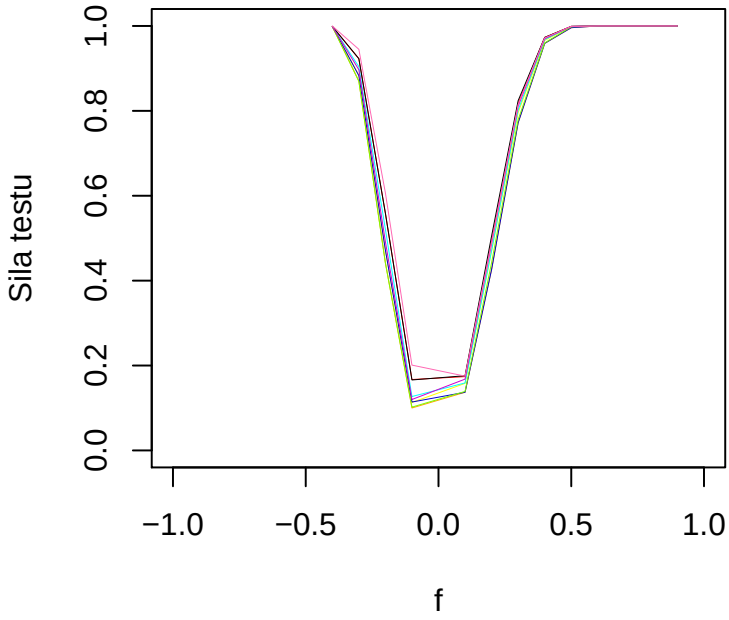
Obr. 6.2: Grafické znázornenie odhadov síl testov pre $p = 0{,}1$



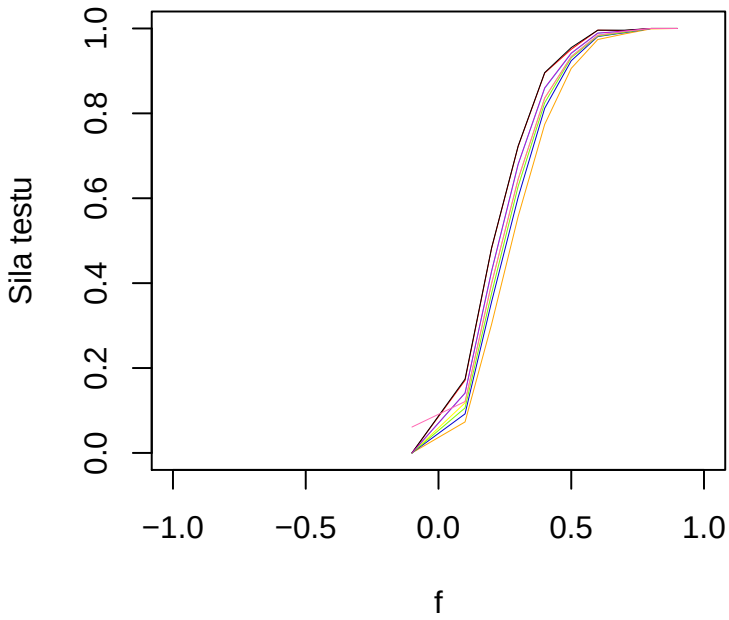
Obr. 6.3: Grafické znázornenie odhadov síl testov pre $p = 0{,}3$



Obr. 6.4: Grafické znázornenie odhadov síl testov pre $p = 0{,}5$



Obr. 6.5: Grafické znázornenie odhadov síl testov pre $p = 0{,}7$



Obr. 6.6: Grafické znázornenie odhadov síl testov pre $p = 0{,}9$

Záver

V tejto práci sme sa venovali testom, ktoré sa používajú pri overovaní, či daná populácia je v Hardyho-Weinbergovej rovnováhe. V prvej kapitole sme sa venovali multinomickému rozdeleniu a to z toho dôvodu, že ak je populácia v Hardyho-Weinbergovej rovnováhe, môžeme rozloženie genotypov, ktoré k nej prislúchajú, popísať trinomickým rozdelením, s pravdepodobnosťami θ^2, $2\theta(1-\theta)$, $(1-\theta)^2$. Neskôr sme popísali niektoré testy a overili oprávnenosť ich použitia pri tomto probléme.
V poslednej kapitole sme sa venovali odhadnutiu hladiny a sily testov. Pri veľkosti populácie $n = 100$, nám ako najvhodnejšie testy pre riešenie daného problému vyšiel modifikovaný χ^2 test. Ako najmenej vhodné, tj. najkonzervatívnejšie a s najnižšou odhadovanou silou testu vyšli exaktný test s prvým spôsobom výpočtu P-hodnoty, χ^2 test s opravou na spojitosť $c = 0{,}5$ a modifikovaný χ^2 test s opravou na spojitosť $c = 0{,}5$. Ako antikonzervatívne sa v niektorých prípadoch javili χ^2 test a test pomerom vierohodnosti. Nesmieme však zabúdať, že tieto výsledky nám vznikli pre $n = 100$, pri iných veľkostiach populácií vyjde odhadovaná hladina testov a sila testov inak.

Zoznam použitej literatúry

[1] Anděl, Jiří. *Základy matematické statistiky.* 2. vydání. Matfyzpress, Praha 978-80-7378-162-0

[2] Ching, Chun Li: *Population Genetics.* University of Chicago Press, 1955

[3] Cramér, H.: *Mathematical Methods of Statistics.* Princeton, Princeton Univ. Press, 1949

[4] Elston, R. C Forthofer, R.: Testing for Hardy-Weinberg equilibrium in small samples. *Biometrics,* 33, 536-542, 1977

[5] Emigh, T. H.: A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics,* 36, 627-642, 1980

[6] Graffelman, J. Camarena, J. M.: Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity,* 65, 77-84, 2008

[7] Wigginton, Janis E. Cutler, David J. Abecasis Conçalo R.: A note on Exact tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics,* 76:887-883, 2005

Zoznam tabuliek

6.1 Odhad hladiny testov 23
6.2 Intervaly spoľahlivosti pravdepodobností I. druhu 23
6.3 Odhad sily exaktného testu s 1. definíciou P-hodnoty 23
6.4 Odhad sily exaktného testu s 2. definíciou P-hodnoty 24
6.5 Odhad sily χ^2 testu 24
6.6 Odhad sily χ^2 testu s opravou na spojitosť $c = 0{,}5$ 24
6.7 Odhad sily χ^2 testu s opravou na spojitosť $c = 0{,}25$ 24
6.8 Odhad sily modifikovaného χ^2 testu 25
6.9 Odhad sily modifikovaného χ^2 testu s opravou na spojitosť $c = 0{,}5$ 25
6.10 Odhad sily modifikovaného χ^2 testu s opravou na spojitosť $c = 0{,}25$ 25
6.11 Odhad sily testu pomerom vierohodnosti 25

Zoznam obrázkov

6.1 De Finettiho diagramy pre $n = 10$, $n = 20$, $n = 50$ 21
6.2 Grafické znázornenie odhadov síl testov pre $p = 0{,}1$ 27
6.3 Grafické znázornenie odhadov síl testov pre $p = 0{,}3$ 27
6.4 Grafické znázornenie odhadov síl testov pre $p = 0{,}5$ 28
6.5 Grafické znázornenie odhadov síl testov pre $p = 0{,}7$ 28
6.6 Grafické znázornenie odhadov síl testov pre $p = 0{,}9$ 29

Zoznam príloh

Na priloženom CD sa nachádzajú nasledujúce prílohy:

Príloha č. 1: HWChisqM.txt–χ^2 test

Príloha č. 2: Lratio.txt– test pomerom vierohodnosti

Príloha č. 3: simulaciahladina.txt– simulácia dát a odhad hladiny testov

Príloha č. 4: simulaciasila.txt– simulácia dát a odhad sily testov

Príloha č. 5: ternary.txt– náčrt de Finettiho diagramov